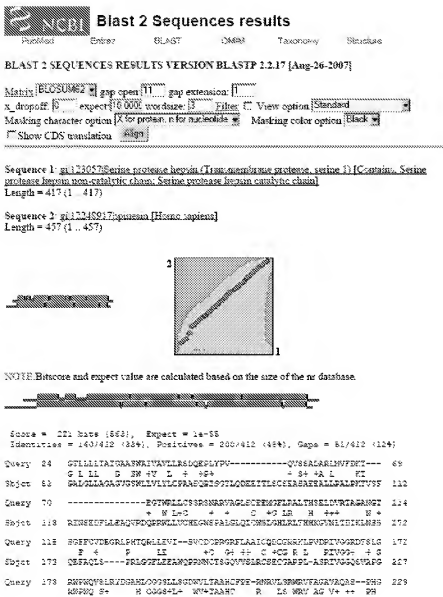


## Appendix 1. BLAST alignments

### A. hepsin vs spinesin:



```

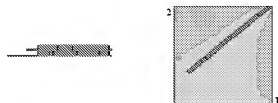
Subject 218 RSPWQASVALGERMTCCGSLAPKKVTAHYNHFFLLBLSTFMRKRLIVEREA/SPHQ 237
Query 230 LQLCTQAVVTHGCLPFFEDHSEKSNQIALYHLSFLFLKXIQPVYLPAGGRLVOK 288
      L Y* ++ H E      + + + + DVAL: L + L ++ + VTEPA Q Q
Subject 288 GAI-VERIIFHFL-----SAGWDMVYGLLRGLRALHFEETVQAVYLPATEQRHFRKF 348
Query 280 ICTVIGKCHYT-YYQQAATLQERKVDISHQVNTADFFVGNDFPWNKALYFEDLGA 348
      C WAGDQVY + + +LQ+ VE+ S +DM + Y + E+H JADE +D EA
Subject 342 POWFSKAWHFRSHIYESHMLGHVVPLESTQLONSCHYSGALCPHMLCATILYGRAM 400
Query 345 TQCTGGGFTVTEDEISRIFRRLQGVNWTGALAQKFGVTKVSDPREWI 400
      CQGGGGFT VC D      NDL G+VSDG CA FVVL EY+P +RI
Subject 401 DGGGFTFLVDFQ-----LGNALVGVHESGALAEHHPQVAKVASTFLKI 448

```

## B. hepsin vs TMPRSS2:

Sequence 1: [gi|33097|Serine protease hepsin \(Transmembrane protease, Serine 1\) \[Contains: Serine protease hepsin non-catalytic chain, Serine protease hepsin catalytic chain\]](#)  
Length = 417 (1...417)

Sequence 2: [gi|46024|Transmembrane protease, serine 2 \[Homo sapiens\]](#)  
[gi|360174|gb:AAD71117.1:AF124451\\_1 non-membrane serine protease 2 \[Homo sapiens\]](#)  
Length = 452 (1...492)



NOTE: Hitscore and expect value are calculated based on the size of the nr database.

Score = 210 bits (634), Expect = 9e-52  
Identities = 117/348 (33%), Positives = 174/348 (50%), Gaps = 17/348 (4%)

Query 68 LMFCKTEITWFLCSFRNARVAGLSCEH/PLAALTHSELVRIAGANGTSQFFCVDE 122  
L ++ +W +C N +D +W+ S + V +G+  
Sbjct 185 LQWSSQKKEWEPVQCDKRENYDPAAGDNGYKNFISQIVDSQSTFENSLN-TSR 216

Query 123 GPLSKTGLLEIVSVDDPPGFLALICQDCG--KHLFVDSIVQVSDTSLSHFQVEL 180  
G + ++L + C + + C QG RINGG G KPTVEL  
Sbjct 217 GRDITVLYEYDA---CSKAVVELACIACFQNLSSDQERIVGSEALFRAHWQVEL 270

Query 181 RIDGAILGGSLSGWVLTAACTFDPNNVLSNNVFAAGVAGASE-WGLQVTKVTV 236  
K+CGSG+++ +N+DAAC + N FAS + Q+ +D V+ V+  
Sbjct 274 HYNVWVQVPSITFENITAAACVKKLNSPHWTAFAQLRQSNFYGACVQVKVIS 338

Query 240 KGLLTFEPFUSEKNSHDLVLESEFLQTEYTOPVLRAGCALVDGKICVTWNT 298

H V +S+ +NDAL+ L EL + --PQLS G L ++C ++WQ V  
Sbjct 336 HENV-----SEKTHNDIALMLNPLTFNDEKVPCLRPQGMQLQEQUNISGWGAT 387

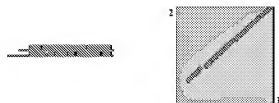
Query 300 QHYGGAGVGLGARVFIISNVYNGADFKVQIKFSEPCMYEFGQIDACQGLSGGTV 358  
+ G+ + VL R+V +I QY V V + S N GAG +G -SG+CGSGG V  
Sbjct 348 EEFQVTEVLMARVLIETQKCHERTVQPLTFAKICAGFLQVFCSCQDQDQDRLV- 448

Query 360 EDHSKTEVPLQGVVSGQTCALAGFQVVTYKIDFRENIFCAIK 406  
S+ + L Q RQ+QD R PQT V F +H+ + +  
Sbjct 447 ---TEKHIMVLIQDTHVIGCAKAYRQVQVTFPFTDNIYQMR 489

# C. hepsin vs TMPRSS3:

Sequence 1: gi|13057|Serine protease hepsin (Transmembrane protease, serine 1) [Contains: Serine protease hepsin non-catalytic chain; Serine protease hepsin catalytic chain]  
Length = 417 (1..417)

Sequence 2: gi|17182182|TMPRSS3 (Homo sapiens)  
Length = 432 (1..432)



NOTE: Bitscore and expect value are calculated based on the size of the m database.

```
Score = 271 bits (844), Expect = 5e-86
Identities = 128/340 (38%), Positives = 179/340 (53%), Gaps = 34/340 (10%)

Query  50  PLVAVVSSADAGLAVFQFTSTWLLCSSEMGAVAGLSCEZNGFLPALI---HSELYD 106
      50+8 + L V D  S N  C  +A +C +HS+ PA+ -LDV
Sbjct 100  PAVAVLSVDRSTLQVLSATGWFACFQMETRAAEFA+PQSVYRATETGPGGLGV 158

Query 107  RTAGAGTSGFFCUDSGRLFTQLLEVISVDCGPGGFLAALCQDQGRMLVDSIVQ 166
      N      +Q L S C G ++ C QG+ L P+VQ
Sbjct 160  VEITEN-----SLELMRSTGGGLSGGVLESLACST+ELTETVQG 203

Query 167  RDIELGPNWQVSLVYDGAELCGGLSGDWLTAHETFPENQVLSKRWYTAGAVAG 216
      + S+ NQWYSS+YD H+CGH+L NVLTAHET+ Y + R+V AG+ E
Sbjct 204  EELAVDSNPAQVSTLQKQVYDGGELIDRWLTAHETFRHETVTH+HWVAGSCWLLS 262

Query 227  PSLQLQVQAVVYHSGFLPFFDEHSEENNDIALVHLSFLPLSTETIQVPLEPAIQALV 286
      L + ++ P      NDIAL+ L PL + ++P+CLP + L
Sbjct 262  PPSLAVAVTIIIEFNHMF-----KNDIATKMLQFLTFSGVGFICLFFDELT 314

Query 287  DSEICTVQHGFT-QYDQAGVLAELVFIISHWVQASSTGCKIKFMFGAGFEGG 345
      + GSG I C D+ +L +A V +I + CH D Y ++ KM CG FEGG
Sbjct 316  SATPLMIGWFTVQMGWQSILLQAEVQVITSTPGNADRAVQGVILMSICAGFEGG 374

Query 346  IDACQDSGFFTVHDSIGATFAGRLCGVSMVICALAGSPVYTVSDFFETFAIK 406
      +D CQDQSGG + + + +M + GIVSG GC EGVITPV + ME+ K
Sbjct 376  VDTQDQSGGFLNYC-----SDMRTVGLVSGDGGSPETPVTYTVSAVLMNIVNK 418
```

#### D. hepsin vs TMPRSS4:

Sequence 1. [gi|130572](#):Serine protease hepsin (Transmembrane protein, chain 1) [Contains: Serine protease hepsin non-catalytic chain; Serine protease hepsin catalytic chain]  
Length = 417 (1 - 417)

Sequence 2: gff1545:940Transmembrane protease, serine 4 isoform 1 [Homo sapiens]  
gff1563980:Q02783:4TMAP5\_HUMAN Transmembrane protease, serine 4 (Membrane-type serine  
protease 2) (MSP-2) (gff158374:gbfAAAF45261:AAFI79224\_1 transmembrane serine protease 3  
[Homo sapiens]) gff1507979:gbfAAH11703\_1 Transmembrane protease, serine 4 [Homo sapiens]  
gff42077743:gbfBAD18749\_1 unnamed protein product [Homo sapiens]  
Length = 437 (1..437)



NOTE Bitscore and expect value are calculated based on the size of the nr database.



Index	Size	Time	Space	Score
Index	215	168	105	100
Index	126/357	105/1	100/357	100/1
Query	50	11/11	11/11	11/11
Subject	100	11/11	11/11	11/11
Query	110	11/11	11/11	11/11
Subject	150	11/11	11/11	11/11
Query	170	11/11	11/11	11/11
Subject	210	11/11	11/11	11/11
Query	230	11/11	11/11	11/11
Subject	270	11/11	11/11	11/11
Query	290	11/11	11/11	11/11
Subject	330	11/11	11/11	11/11
Query	350	11/11	11/11	11/11
Subject	390	11/11	11/11	11/11
Query	410	11/11	11/11	11/11
Subject	450	11/11	11/11	11/11

# E. hepsin vs enteropeptidase:

Sequence 1. [gi122677](#) Serine protease hepsin (Transmembrane protease, serine 1) [Contains: Serine protease hepsin non-catalytic chain, Serine protease hepsin catalytic chain]  
Length = 417 (1 - 417)

Sequence 2. [gi6909361](#) Enteropeptidase [Homo sapiens]  
Length = 1019 (1 - 1019)



NOTE: Bitscore and expect value are calculated based on the size of the nr database.

```

Score = 111 bits (536), Expect = 1e-82
Identities = 106/256 (41%), Positives = 145/256 (56%), Gaps = 17/256 (6%)

Query 151 QDCGRRLKLV---PIVGGDTSLGRNPGVSLRIDIAHLQPSLLSGDNYLTAHCFPE 207
          + CG++ D +IVGG + G NFN V L Y S LQ SL+ DN+++AHC
Sbjct 770 KDCGRLLAAGDITPIVGGANALGNPHVSLIYQWLLDASINVSQWIVAAHCVVG 829

Query 208 PNRVLSRNQVTAAYAGA---SPSLQLGVQANVTAGGYLSTFDPNHEENNDIALVSL 264
          RH S+M G ++ SP + + +V + Y N NDA--HL
Sbjct 990 RNLSPKNTAILQNRKTHLTPTVPRLLDIIVNPHI-----NSRPNNDIAHGLE 998

Query 266 SPLESLTYITQVPLPAGQALVGRKLTITVQNGTQITQAGVLEARVLEISNDVNG 324
          + T-YIQ+ILP Q G+ G++ GNG Y G R -LQA VP++SH+ C
Sbjct 994 FHTVITDTIQICLFEINQFFFFRRHSLAAGGTVTQSTARKLQEAIVLLNENQ~ 942

Query 328 ADFYGNQIKTFKFCAGYFEGGIDAGQGDSDGGFVCELSISRTPRNLOSIVNATGALA 394
          L M CAGY EGGD+QGGEGGP +D+++ SW L G+ S+G CAL

Sbjct 943 QMFPEINITENKICAGYEGGIDAGQGDSDGGFVCELSISRTPRNLOSIVNATGALA 998

Query 388 QKQGVTKVSEFEMI 400
          +PTV +S F EWE
Sbjct 998 NRPSTARVSEFEMI 1014

```

## F. hepsin vs MSPL:

Sequence 1: gi123057Serine protease hepsin (Tricusmenus proteus, serine 1) [Contains Serine protease hepsin non-catalytic chain: Serine protease hepsin catalytic chain]  
Length = 417 (1..417)

Sequence 2: gi11676663transmembrane protease, serine 12 (Homo sapiens)  
Length = 367 (1..367)



NOTE Bitscore and expect value are calculated based on the size of the nr database.

```

Score = 231 bits (888), Expect = 1e-58
Identities = 124/372 (36%), Positives = 196/372 (52%), Gaps = 29/372 (7%)

Query 38  IVNVLLEDEPEFLVWVSSADAMVFWTEGTWELLOSRSNNAKGLSCENGGFLRA 58
      +V L+ED+ L V+ + L ++ + S +SS N + +C++GF A
Sbjct 214  VVCELEADE--LQVAFWDEKLLRTSOSRQNLPTOSNNWDSYKEDQDZ/EEEA 274

Query 58  LTHSELDVITAGANSTSEFFVDEHRLSTQRLLEVIEWDCEPPFELAIQDQGRKIL 168
      +E+ R F L + + E + +CF F++ C GG R +
Sbjct 272  KATTEVAHD-----FANSFSLRNWTLCEELHSEDESPRYISLQCHGCSLAN 322

Query 188  FVRIVQSDTSLGRMPROVLENDGRLQGLLGGWVLIANHC--FFRNVLSEWV 217
      RIVDS S +MPPQVIL + R+DQ+L+ VLIANHC F R +VL M+V
Sbjct 328  -TQIVVGLALASDEKQWQVSLHFTITRIGSLIDAGVLIANHTFFVIRENVLAGKV 381

Query 216  FAFVAAQALPHGL--GLVYAAVYVNGVYLFPRFHEENRUDALVHSSPLTEVIG 275
      +AG ++ H L + ++ + Y EE+ DIAL+ LS FL L+ +I F
Sbjct 382  YAGT---GULAGLPEASTAETIINENY-----TQRETOYDIALMRSLKPLTSANHP 432

Query 276  VILPARGQALVDGKICVTVGWNTQYTPQDQG-VLQKARVFIENQVWVADPFVQVNP 324
      CLS QQ + C +EG+2 I+ + LHE +V +2 CH Y + 2
Sbjct 433  ACLHNGQVITSINETCNLIGFQVTPETDQKTSFLRELVQVNLLEDFPQNDVLYVDSVLFP 492

Query 386  IGFVATYSEGGIDAGQDQDQDQDPTKEDSLRTFVHMLQGVVSWVQALAQKGVVTVH 394
      +H CAG QQ D+QDQDQDQD VCK + PR L Q+ SWGTG KQGVTVH+
Sbjct 493  RHTAGAGLQAGDQDQDQDQDQDQDQDQDQDQDQDQDQDQDQDQDQDQDQDQDQDQD 548

Query 388  DEFEVTSKAIT 406
      + VE+
Sbjct 549  EVLEWTVSRNEZ 560
  
```